Heterogeneous Face Recognition: Matching NIR to Visible Light Images

Brendan Klare and Anil K. Jain[1]
Department of Computer Science and Engineering
Michigan State University
East Lansing, Michigan, U.S.A.
{klarebre, jain}@cse.msu.edu

***Abstract*—Matching near-infrared (NIR) face images to visible light (VIS) face images offers a robust approach to face recognition with unconstrained illumination. In this paper we propose a novel method of heterogeneous face recognition that uses a common feature-based representation for both NIR images as well as VIS images. Linear discriminant analysis is performed on a collection of random subspaces to learn discriminative projections. NIR and VIS images are matched (i) directly using the random subspace projections, and (ii) using sparse representation classification. Experimental results demonstrate the effectiveness of the proposed approach for matching NIR and VIS face images.**

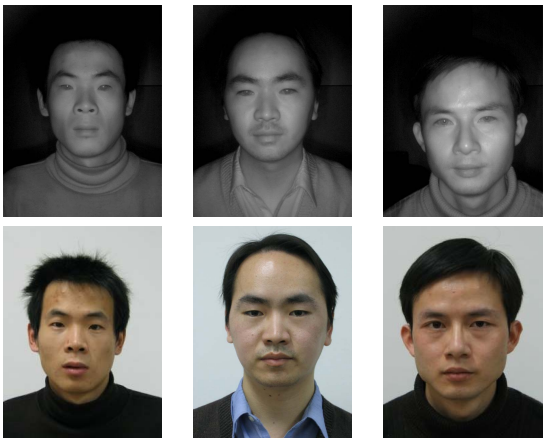
***Index Terms*—Face recognition; near infrared; feature-based; random subspaces; spare representation;**

I. Introduction

The degradation in face recognition performance due to unconstrained illumination is well documented [12]. A solution to face recognition in the presence of varying illumination is to acquire face images beyond the visible spectrum (VIS) Near-infrared digital images (NIR) measure the presence of electromagnetic radiation just beyond the visible light range (specifically, optical radiation with wavelengths between .7μm and 1.4μm). Images acquired in the near-infrared spectrum are close enough to the visible light spectrum to still capture the structure of the face, yet far enough removed to not change the facial appearance due to visible light illumination changes. An example of NIR and VIS images is shown in Figure 1.

Many face recognition scenarios involve matching probe NIR images against a previously acquired visible gallery database, such as mugshots or passport photos. The ability to match NIR probe images to the gallery VIS images is of significant importance in scenarios that require query face images to be acquired in poor illumination conditions, such as nighttime. However, in matching NIR images to VIS face images, difficulties are introduced from matching across image modalities (referred to as heterogeneous face recognition). In this paper we present a robust solution to the heterogeneous face recognition problem of matching between NIR and VIS face images.

A few methods of matching NIR and VIS face images have been proposed. Yi et al. used canonical correlation analysis [16] to learn the similarity between NIR and VIS images by training on NIR/VIS image pairs of the same subject. Wang et al. matched NIR images to VIS images directly by converting NIR images to synthetic VIS images [13]. Chen et al. used local linear embedding with a dictionary of corresponding NIR/VIS face pairs to convert NIR images to synthetic VIS images for matching [3]. Liao et al. used local binary patterns (LBP) to describe both NIR and VIS images, and performed matching between the LBP representations of the two images using LDA [10].



(a) (b) (c)

Fig. 1. Examples of near infrared images (top row) and visible light images (bottom row) for the same subjects. Images shown are from the CASIA HFB Database [9]

Our approach to matching NIR and VIS face images is based on the method of Liao et al. [10], but offers the following improvements: (i) using histograms of oriented gradients (HOG) feature descriptors in addition to LBP descriptors to extract more salient information regarding the structure of the face, (ii) learning an ensemble of discriminant projections to improve the generalization and better handle the high dimensionality of the feature representation, (iii) incorporating sparse representation classification, which has shown some merit in face recognition [15], and (iv) using a commercial face recognition engine to define the baseline performance.

[1]Anil K. Jain's research was partially supported by World Class University (WCU) program through the National Research Foundation of Korea funded by the Ministry of Education, Science and Technology (R31-2008-000-10008-0).

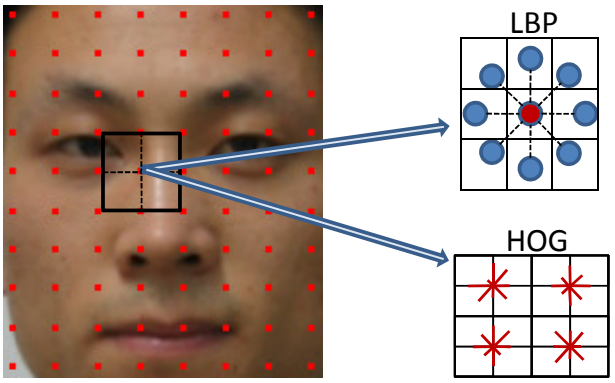


Fig. 2. Feature descriptors are sampled at equally spaced points on the face.

II. Matching Algorithm

This section will describe the proposed method for matching near-infrared to visible light images. The method uses a set of n training subjects to learn discriminative feature projections. Training subject $i = (1, \ldots, n)$ will have n_i^N NIR images, which are denoted $I_{i,j}^N$, $(1 \leq j \leq n_i^N)$. Similarly, the i-th training subject will have n_i^V VIS images, which are denoted $I_{i,j}^V$, $(1 \leq j \leq n_i^V)$.

A. Face Image Representation

Each NIR and VIS face image is described using HOG and uniform LBP image feature descriptors in order to compensate for the difference between the two image modalities while still preserving information that describes the structure of the face. The effectiveness of LBP descriptors for representing both NIR and VIS images has already been documented [10]. HOG descriptors (more commonly known as the feature descriptor in the SIFT framework [11]) was successfully used for matching between face sketches and photographs [7], [8], and similar strengths were witnessed in representing NIR and VIS images.

To represent a face image using image features, each face image is first normalized by (i) rotating the image so that the angle between eyes is 0°; (ii) scaling the image to an interocular pixel distance of 100 pixels; (iii) cropping the image to a size of 250x200 pixels. The boxes on each face in Figure 3 show the cropped region for each face. Tight cropping was needed in order to remove hair from the NIR images. This process if fully automated.

For each normalized face image, we next sample a set of P uniformly spaced patches of size s x s pixels from the face and extract the image features (LBP and HOG) from each of these patches using the same method described for sketch matching [7]. Figure 2 illustrates this feature extraction process. For a face image I, LBP feature extraction is denoted by the function $f_L(I)$, which results in a set of P LBP feature vectors $l_p \in \mathbb{R}^{59}$ for each image patch, i.e. $f_L(I) = (l_1, l_2, \ldots, l_P)$. Similarly, the HOG feature extraction is denoted by the function $f_G(I) = (g_1, g_2, \ldots, g_P)$, $g_p \in \mathbb{R}^{128}$. Combining these two features, we describe a face image by an ordered set of $2P$ features vectors $f(I) = (f_L(I), f_G(I))$. For the j-th NIR image from subject i, we denote the ordered set of feature descriptors as $X_{i,j}^N = f(I_{i,j}^N)$. The p-th feature vector from $X_{i,j}^N$ is denoted as $X_{i,j}^N(p)$, where $X_{i,j}^N(p) \in \mathbb{R}^{59}$ if $1 \leq p \leq P$, and $X_{i,j}^N(p) \in \mathbb{R}^{128}$ if $P < p \leq 2P$. The j-th VIS image from subject i is denoted $X_{i,j}^V = f(I_{i,j}^V)$.

B. Feature-based Random Subspaces

Given the feature-based representation $f(I)$ of near infrared and visible light images, we seek to reduce the feature dimensionality by learning discriminative projections via linear discriminant analysis (LDA). However, when used in face recognition, LDA methods are: (i) prone to overfitting due to the fact that few training samples generally exist for each class/subject, and (ii) limited in the number of features that can be extracted by the total number of classes. To compensate for these deficiencies, we use an ensemble classifier trained on random subspaces [5]. This approach was successfully applied to pixel intensities [14] for matching between visible light face images.

Our random subspace method using the feature-based representation of NIR and VIS face images is performed as follows. For each iteration $k = (1, \ldots, K)$, we randomly sample (without replacement) α feature vectors, where $1 \leq \alpha < 2P$. For each set of feature vectors $X_{i,j}^N$ and $X_{i,j}^V$, the α randomly selected vectors are concatenated into a single vector $x_{i,j}^N(k) \in \mathbb{R}^d$ or $x_{i,j}^V(k) \in \mathbb{R}^d$, where the size of d is in the range $[59 \cdot \alpha, 128 \cdot \alpha]$ based on how many vectors are selected from the LBP and HOG feature vectors. The mean class vector for each subject at iteration k is then built using both the NIR and VIS images together as

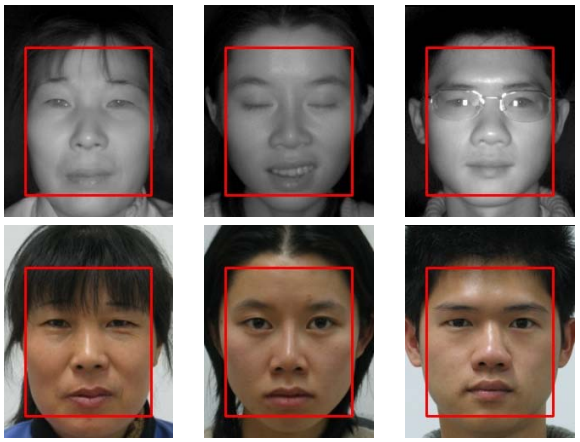
$$\mu_i^{(k)} = \frac{1}{n_i^V + n_i^N} \left(\sum_{j=1}^{n_i^V} x_{i,j}^V(k) + \sum_{j=1}^{n_i^N} x_{i,j}^N(k) \right) \tag{1}$$

These mean vectors are used to construct both the between-class and within-class scatter matrices. The between-class scatter matrix $S_B^{(k)}$ for iteration k is computed as $S_B^{(k)} = \sum_{i=1}^{n} (\mu_i^{(k)} - \mu^{(k)})(\mu_i^{(k)} - \mu^{(k)})^{\mathrm{T}}$ where $\mu^{(k)} = \frac{1}{n} \sum_{i=1}^{n} \mu_i^{(k)}$. The within-class scatter matrix is computed as

$$\begin{aligned} S_i^{(k)} = &\sum_{j=1}^{n_i^V} (x_{i,j}^V(k) - \mu_i^{(k)})(x_{i,j}^V(k) - \mu_i^{(k)})^{\mathrm{T}} \\ &+ \sum_{j=1}^{n_i^N} (x_{i,j}^N(k) - \mu_i^{(k)})(x_{i,j}^N(k) - \mu_i^{(k)})^{\mathrm{T}} \end{aligned} \tag{2}$$

The total within-class scatter is $S_W^{(k)} = \sum_{i=1}^{n} S_i^{(k)}$.

The final step in each iteration of the feature-based random subspace method is to compute the matrix of eigenvectors $V^{(k)}$ that satisfies the generalized eigenvalue problem $S_B^{(k)} V^{(k)} = \lambda^{(k)} S_W^{(k)} V^{(k)}$. Because $d > n$, which causes both scatter matrices to be singular, the dimensionality of the training vectors $x_{i,j}^V(k)$ and $x_{i,j}^N(k)$ is reduced using the PCA projection matrix $W^{(k)}$ prior to computing the LDA scatter matrices [2]. This requires that all feature vectors $x_{i,j}^N$ be mean centered.



(a) (b) (c)

Fig. 3. Instances in which (a) *FE-SR* succeeded but *FV* failed, (b) *FV* succeeded but *FE-SR* failed, and (c) both *FE-SR* and *FV* failed, but *NNSR+FV* succeeded. Boxes indicate the cropped region used for face recognition.

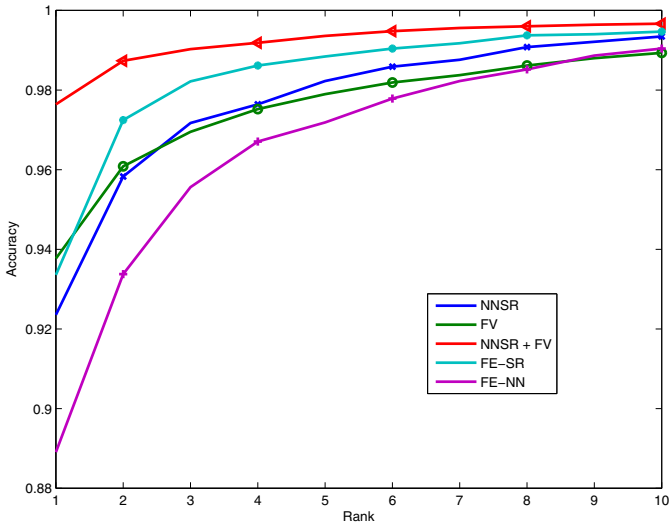


Fig. 4. CMC plots of the performance of the proposed methods and a commercial face recognition engine in an identification scenario.

At the end of each iteration k, we learn a discriminative projection matrix $\Phi^{(k)} = (V^{(k)})^{\mathrm{T}}(W^{(k)})^{\mathrm{T}}$. Thus, the new representation of the α selected feature vectors is $y^N_{i,j}(k) = \Phi^{(k)} x^N_{i,j}(k)$, where $y^N_{i,j}(k) \in \mathbb{R}^{d'}, d' \leq (n-1)$

C. Matching

Once our training phase is complete, we are able to perform matching on face images of subjects that were not present in the training set. Each VIS gallery image or NIR probe image I is first converted into their feature based representation $X = f(I)$. For each of the K feature-based random subspaces learned in the training phase, we select and concatenate the same α feature vectors choosen at the corresponding iteration to generate the vector $x(k)$. After computing each of the K projected representations $y(k) = \Phi^{(k)} x(k)$, $(k = 1 \ldots K)$, we can concatenate each of these d'-dimensional vectors into $Y = [y(1)^{\mathrm{T}}\ y(2)^{\mathrm{T}}\ \cdots\ y(K)^{\mathrm{T}}]^{\mathrm{T}}, Y \in \mathbb{R}^{K \cdot d'}$.

a) Nearest Neighbor Matching: Using the final representation Y from a face image I, nearest neighbor matching can be performed to find the identity of a probe NIR image. For example, suppose we are given a probe NIR image I^N of an unknown subject, and n' VIS gallery images $I^V_1, I^V_2, \ldots, I^V_{n'}$. We first represent each image using its random subspace representation ($I^N \rightarrow Y^N$, and $I^V_i \rightarrow Y^V_i$), and choose the gallery subject that has the minimal Euclidean distance from the probe image: $\text{Identity}(I^N) = \underset{i}{\operatorname{argmin}} \|Y^V_i - Y^N\|_2$.

b) Sparse Representation Matching: Wright et al. have demonstrated [15] the effectiveness of performing face recognition using sparse representation. The same approach can be employed on our feature ensembles. If we let A be a matrix whose columns are the feature representations of each gallery image and y be the probe image feature representation, then by solving

$$x = \underset{x}{\operatorname{argmin}} \|Ax - y\|_2^2 + \lambda \|x\|_1 \qquad (3)$$

the vector x, ideally, will contain non-zero coefficients for only those gallery vectors that belong to the same subject as the probe.

We use a slightly different approach to recovering the identity of a probe image using sparse classification than the method by Wright et al. [15]. For each ensemble $k = (1, \ldots, K)$, we create the gallery matrix $A^{(k)}$ whose columns contain the vectors $y^V(k)$ for each VIS gallery subject. The vector $x^{(k)}$ is recovered by solving Eq. 3, where A is substituted with $A^{(k)}$, and y is substituted with the probe feature vector $y^N(k)$. Eq. 3 is solved using Kim et al.'s interior-point method [6], which is well suited for large scale data.

The final step of our method of sparse ensemble recognition is to create an identity vector $x = \sum_{k=1}^{K} x^{(k)}$ that will accumulate all the non-zero coefficients. The component i that has the maximum value in x is used to indicate the identity of the i-th entry in the gallery.

III. Experimental Results

We evaluated our proposed method of matching near infrared and visible light images on the same images used in previously [10], [4]. This dataset consists of 3,002 NIR and 2,095 VIS images from 202 subjects[1]. We selected 102 subjects to use in training and the remaining 100 subjects for testing. We performed our training and testing five times, each time using a different random split of training and testing subjects. The following parameter values were used: $K = 100$, $P = 720$, $s = 16$, $\alpha = 30$, and $\lambda = .95$

We compared the performance of our method to Cognitec's FaceVACS [1] commercial face recognition engine. The performance of FaceVACS in matching NIR to VIS images demonstrates that baselines used in the literature are very pessimistic.

[1]Portions of this dataset are available for download at: http://www.cbsr.ia.ac.cn/english/Databases.asp

TABLE I
MATCHING PERFORMANCES IN A VERIFICATION SCENARIO. LISTED ARE THE TRUE POSITIVE RATES AT DIFFERENT FALSE POSITIVE (FP) RATES.

FP Rate	0.1 %	1.0 %	10.0%
FE-NN	48.78 ± 3.87	73.49 ± 3.72	95.94 ± 0.56
FE-SR	77.56 ± 2.96	94.04 ± 1.49	99.63 ± 0.21
NNSR	79.05 ± 4.48	91.37 ± 1.99	98.15 ± 0.46
FV	85.62 ± 2.17	93.80 ± 0.65	98.24 ± 0.07
NNSR + FV	93.45 ± 0.96	97.06 ± 0.39	99.41 ± 0.17

Figure 4 shows a plot of the recognition accuracy of our proposed methods, averaged over the 5 training/testing splits (the standard deviation is ~2%). The feature ensemble using the nearest neighbor matching is denoted as *FE-NN*, the feature ensemble using sparse representation is denoted as *FE-SR*, the sum of score fusion of the two methods is denoted as *NNSR*, FaceVACS is denoted as *FV*, and the sum of score fusion of *NNSR* with *FV* is denoted as *NNSR+FV*.

These results demonstrate the effectiveness of the proposed approach. It is seen that the sparse representation classification performs better than the nearest neighbor approach. One of the most intriguing results is that the state-of-the-art face recognition engine (FaceVACS) already performs very well here. Fusing FaceVACS with our feature ensemble recognition yields extremely high accuracy (97.6% at Rank-1).

Table I lists the performances from the proposed methods in a verification scenario. The results are comparative to previous methods for matching NIR and VIS face images. Using 150 subjects for training and 52 testing subjects, Liao et al. [10] achieved a verification rate of 87.5% at a false positive rate of 1%. On the same dataset using 102 training subjects and 100 testing subjects, Table I shows that at a false positive rate of 1%, the following average verification rates were observed: FE-SR 94.04%, FV 93.80%, and NNSR+FV 97.06%. Though these higher performances where generated using a fewer training subjects, it should be noted that the sparse representation method currently does not scale well to large galleries.

In terms of which feature descriptor (HOG or LBP) is more informative, the average Rank-1 accuracy for NN matching using only LBP feature descriptors was 82.5%, while HOG feature descriptors achieved 88.8% accuracy. Instances of successful and unsuccessful matchings can be found in each column in Figure 3, where: (a) *FE-SR* succeeded but *FV* failed, (b) *FV* succeeded but *FE-SR* failed, and (c) both *FE-SR* and *FV* failed, but *NNSR+FV* succeeded.

IV. CONCLUSIONS

We have presented a new approach to matching NIR and VIS face images. Recognition results show that robust solution is available for matching NIR and VIS face images, which should be deployed in face recognition systems that acquire NIR query images in environments with unconstrained illumination. Comparing the performance of the proposed method to a commercial face matcher demonstrated that the baseline for this face recognition is much higher than previously reported. Future work will involve improving the sparse representation matcher to handle larger gallery sizes.

REFERENCES

[1] FaceVACS Software Developer Kit, Cognitec Systems GmbH, http://www.cognitec-systems.de.

[2] P. Belhumeur, J. Hespanha, and D. Kriegman. Eigenfaces vs. fisherfaces: recognition using class specific linear projection. *IEEE Trans. PAMI*, 19(7):711–720, 1997.

[3] J. Chen et al. Learning mappings for face synthesis from near infrared to visual light images. In *Proc. of CVPR*, pages 156–163, June 2009.

[4] R. Chu, S. Liao, and L. Zhang. Illumination invariant face recognition using near-infrared images. *IEEE Trans. Pattern Analysis & Machine Intelligence*, 29(4):627–639, 2007. Senior Member-Li, Stan Z.

[5] T. K. Ho. The random subspace method for constructing decision forests. *IEEE Trans. PAMI*, 20(8):832–844, Aug 1998.

[6] S.-J. Kim, K. Koh, M. Lustig, S. Boyd, and D. Gorinevsky. An interior-point method for large-scale L1-regularized least squares. *IEEE Journ. Selected Topics in Signal Processing*, 1(4):606–617, Dec. 2007.

[7] B. Klare and A. K. Jain. Sketch to photo matching: A feature-based approach. In *Proc. SPIE Conf. on Biometric Technology for Human Identification*, 2010.

[8] B. Klare, Z. Li, and A. Jain. Matching forensic sketches to mugshot photos. In *MSU Technical Report, MSU-CSE-10-3*, 2010.

[9] S. Li, Z. Lei, and M. Ao. The hfb face database for heterogeneous face biometrics research. In *Proc. of IEEE Conference on Computer Vision & Pattern Recognition Workshops*, 2009.

[10] S. Liao, D. Yi, Z. Lei, R. Qin, and S. Li. Heterogeneous face recognition from local structures of normalized appearance. In *Proc. 3rd ICB*, 2009.

[11] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.

[12] PJ Phillips et al. FRVT 2006 and ICE 2006 large-scale results. In *NISTIR 7408*, 2007.

[13] R. Wang, J. Yang, D. Yi, and S. Li. An analysis-by-synthesis method for heterogeneous face biometrics. In *Proc. 3rd ICB*, 2009.

[14] X. Wang and X. Tang. Random sampling LDA for face recognition. *Proc. of CVPR*, 2004.

[15] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *IEEE Trans. PAMI*, 31(2):210–227, 2009.

[16] D. Yi, R. Liu, R. Chu, Z. Lei, and S. Li. Face matching between near infrared and visible light images. In *Proc. 2nd ICB*, 2007.